

## 9. IMPUTATION

### Introduction

Imputation is a collection of methods to deal with missing or inconsistent data in survey responses. For example:

- A survey may ask each respondent to report sales, inventory, and purchase figures, but one company fills in only the sales and inventory blocks on the form. Thus, for that ID, we have two reported data values (sales and inventory) and one missing data value (purchases).
- The respondent may report data for sales, inventory, and purchases, but the value for sales is five times larger than the inventory figure. In previous statistical periods, the company ratio has been about 2:1. How should this inconsistency be treated?
- A keyer may transpose numbers, resulting in values for detail data items that do not equal the total.

One way to deal with inconsistent or missing data is to “impute” a value. Essentially, imputation is an “educated guess” for adjusting such survey data problems. Some adjustments may be simple data corrections that can be entered by examining each data item that failed an edit. Others might require more elaborate techniques to compute a value that meets survey specifications for statistical confidence levels.

The StEPS offers two types of imputation – Simple Imputation and General Imputation.

### Simple Imputation (SIMP)

**Simple imputation (SIMP)** allows you to make minor corrections to a case based upon the values of other data items within that same case. This module sets the data flag for the corrected value to “E,” and the imputed data item is treated as reported. There are two types of Simple Imputation – Free-Form and Balance Complex.

#### Free-Form Simple Imputation

“Free-form” imputation allows an analyst to specify a formula to make simple corrections to data. The formula must specify a test or condition under which the data item should be changed and what to do to change it.

For example:

The survey has an Item named ECYN, which asks “Did your company have any e-commerce revenue last quarter?” The answer is a pair of check boxes for reporting either

“Yes” (value of “1”) or “No” (value of “0”). The next Item is EREV, which asks for the value of the revenue. If ECYN = 0, EREV should be = blank. A respondent checks “No” for ECYN, but gives a value in EREV. A free-form formula might be set up to specify that if EREV is not equal to blank, ECYN should be set to “1.” The changed value would be given a data flag of “E.”

### **Balance Complex Imputation**

“Balance complex,” imputation estimates a value or values within a complex of related items for a given ID. Several data values might be adjusted to ensure that the individual items add up to the expected total. Again, if the confidence level for these adjustments is high, Simple Imputation can be used and the adjusted values treated as if they had been reported.

For example:

Your survey collects four quarterly payroll values ( $P_1$ ,  $P_2$ ,  $P_3$ , and  $P_4$ ) plus a final annual total ( $P_T$ ). If any one of the four quarterly values is missing, you could perform a Simple Imputation, using the other four reported values, to produce a value to replace this missing figure by adding the three detailed data items you have and subtracting the sum from the total [i.e., if the third quarter value is missing, the equation would be:  $P_3 = P_T - (P_1 + P_2 + P_4)$ .] The computed value would be given a data flag of “E.”

### **General Imputation (GIMP)**

**General Imputation (GIMP)** imputes values for items that:

- Fail an edit test and are marked for imputation in the Item Data Dictionary
- Are marked for imputation by an analyst
- Are involved in a failed balance complex test

Again, there are two types of imputation in this module:

- Item imputation, which imputes or adjusts the value of a specific item
- Balance complex, which imputes or adjusts the items in a set so that the details will add to a corresponding total and, if required, to a defined subtotal.

For each of these two basic types, StEPS offers a variety of statistical methods to allow you to tailor the imputation parameters to meet your survey requirements. Data values inserted or adjusted by a General Imputation method are flagged to show the method used for imputation and will be included when calculating the imputation rate for the survey. (See sections 9.3 and 9.4 for a detailed description of these methods and the associated flags.)

## Comparison of Simple vs. General Imputation

The decision about when to use the SIMP or the GIMP is more than the methods available. (Both modules offer item imputation and multi-item balance complex imputation.) The table below compares the characteristics of the two imputation types.

Simple Imputation (SIMP)	General Imputation (GIMP)
Is self-contained and does not use the edit tests. SGMP specifications define the test or condition under which a data item should be changed and how to change data.	Requires the Edit specification to have the imputation action flag (IMPACT) set to impute the item if the edit tests fail. The GIMP specs only define the methods (the “how”) to change the data. They do not define which data to change.
Item data flag is set to “E” and the data is treated as if it were reported (data flag = “R”)	Item data flag is set to show method used to impute value and the data is treated as imputed
Item not included when Imputation Rate calculated	Item included in Imputation Rate

The flowchart shown as Figure 9.1 illustrates the flow of Survey processing as it might occur in a night run using both SIMP and GIMP modules. (For simplicity, this overview does not show the Derive and Adjust modules.)

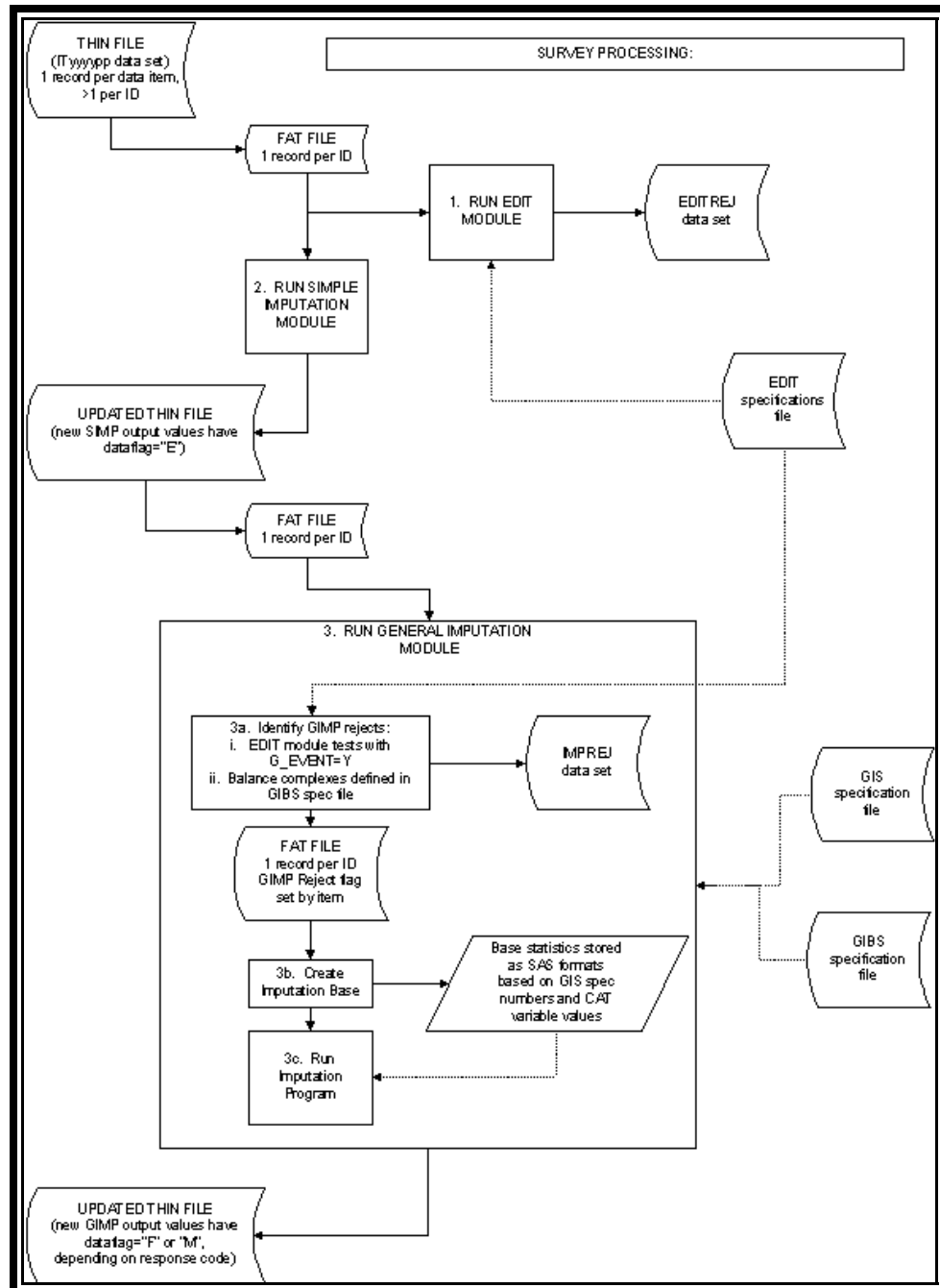


Figure 9.1 Imputation Processing Flow

## Organization of This Chapter

This chapter is organized, as follows:

Section 9.1	Glossary
Section 9.2	Simple Imputation methods available in StEPS and using the StEPS to create Simple Imputation specifications.
Section 9.3	General Imputation methods available in StEPS and a discussion of the mathematical theory associated with each method.
Section 9.4	Creating a General Imputation specification using the StEPS screens.
Section 9.5	Specifying Relative Stat Periods for Imputation.
Section 9.6	Imputation for Roster Items